Extending the Dual-system Estimation for the Census of Agriculture

Habtamu Benecha[*] Luca Sartore[*†] Grace Yoon[*] Bruce A. Craig[*‡]
Denise A. Abreu[*] Linda J. Young[*]

Abstract

USDA's National Agricultural Statistics Service (NASS) conducts the Census of Agriculture every five years. The Census is the leading source of information on U.S. agriculture, providing characteristics of farms and the people who operate them. However, the Census Mail List (CML) is incomplete. That is, the CML does not contain all operations that are farms, and some records on the CML are not farms. To quantify the incompleteness in the CML, NASS uses the June Area Survey (JAS), which is based on an area frame. Census weights are adjusted by applying a capture-recapture method that accounts for undercoverage, non-response, and misclassification. Historically, only JAS data and CML records that are linked to the JAS are used for producing these adjustment weights. This paper proposes an alternative capture-recapture approach that utilizes all the available JAS & Census information for the estimation of Census weights. Results from simulation studies and an application of the method to data from the 2017 Census of Agriculture are presented.

Key Words: Capture-recapture, Coverage, Non-response, Misclassification, List frame, Area frame.

1. Introduction

USDA's National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture every five years (in years ending in 2 and 7). The Census is the only source of uniform comprehensive agricultural information for every state and county in the United States (U.S.), and counts U.S. farms, ranches, and the people who operate them. By definition, a farm is any agricultural operation from which $1,000 or more of agricultural products were produced and sold, or normally would have been sold, during the Census year (USDA National Agricultural Statistics Service, 2019).

As a frame for the Census, the Census Mail List (CML) is a list of all known farms and potential farms. It is constructed by updating the list of farm operations in previous censuses and by incorporating information from annually-collected sources. During the Census, a questionnaire is sent to each operation on the CML. Because the CML is incomplete, NASS uses the June Area Survey (JAS) to quantify the undercoverage in the CML and produce adjusted Census estimates. The JAS is based on an area frame that covers all land in the continental U.S. with every acre of land having a known probability of selection. Prior to distributing the Census questionnaires, CML and JAS records are linked. The subset of JAS records that are not linked to the CML are called Not-on-the-Mail-List (NML) records. If a record in the NML domain is deemed a farm during the Census, it is recorded as NML farm. The NML records and their NML farm classification are used to measure coverage associated with the Census (USDA National Agricultural Statistics Service, 2019).

[*] United States Department of Agriculture, National Agricultural Statistics Service, 1400 Independence Avenue SW, Washington, DC 20250

[†] National Institute of Statistical Sciences, 1750 K Street NW Suite 1100, Washington, DC 20006

[‡] Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, West Lafayette, IN 47907

In addition to CML undercoverage, adjustments need to be made for non-response, and misclassification of farms. Misclassification of farms occurs when farms are counted as non-farms or non-farms are counted as farms. Consequently, NASS employed a capture-recapture method for producing adjusted estimates for the 2012 & 2017 Censuses (Young et al., 2017, 2013). This involved fitting various logistic regression models to different subsets of the JAS & linked CML records. These models combined produce adjustment weights for undercoverage, non-response, and misclassification for responding Census records. While this method allows for the estimation of an adjusted weight for each record in the Census, only a small fraction of Census records are used in the estimation of model parameters. In addition, separate models need to be fitted for the estimation of adjustment weights for undercoverage, non-response, misclassification overcounting (i.e., when a non-farm is counted as a farm) and misclassification undercounting (i.e., when a farm is counted as a non-farm).

In this paper, a unified Census modeling framework is developed for improving the estimates and the estimation process by utilizing all the available Census & JAS data. All parameters are simultaneously estimated from one model. The proposed approach extends the methods discussed in Alho (1990) under the assumption of correct classification of the units listed in the Census and JAS (see also Huggins 1989; Alho et al. 1993; Alho 1994). The coverage probabilities for both the CML and JAS are estimated together with the probability of responding to the Census questionnaire given that the record was listed in the CML. The inverse of the inclusion probability (provided by the product of the CML coverage and conditional response probabilities) is used to compute the weights associated with each observed Census record.

The 2017 Census data allows for a fair comparison of the results produced from the existing methodology applied in 2017 and the proposed approach. From several diagnostics computed on the Census data, various modeling strategies have been investigated (such as fitting the model by state) leading to the computation of more accurate estimates. In addition, the model penalizes high weights at the record-level to avoid unrealistically high estimates at the county level. Simulation studies are conducted and the method is applied to the 2017 Census to estimate the total number of farms and total land operated by farms at the state and national levels.

This paper is structured as follows: a brief background of the JAS, the Census, and the CML is presented in Section 2. Section 3 describes the development of the likelihood (see Section 3.1), adjustment for JAS sampling probabilities (see Section 3.2), the penalty to control extreme weights (see Section 3.3), and the estimation of the model parameters and population totals (in Section 3.4). Results from simulation studies are presented in Section 4, and a case study that focuses on the 2017 U.S. Census of Agriculture is provided in Section 5. Concluding remarks are discussed in Section 6.

2. The June Area Survey, the Census and the Census Mail List

2.1 The June Area Survey

The June Area Survey is one of the largest annual surveys conducted by NASS and is used to collect information about U.S. crops, livestock, grain storage capacity, and farm sizes and types (Lamas et al., 2010). The JAS uses a stratified area frame that covers all land in the U.S. except Alaska. The strata, which are based on the percent of land devoted to agriculture, are further divided into substrata by grouping areas that are agriculturally similar. Within each substratum, the land is divided into primary sampling units (PSUs). A sample of PSUs is selected and smaller, similar-sized segments of land, which are each about a square mile in area, are sampled from the selected PSUs to be fully enumerated. Before the survey, all tracts of land within selected segments are screened and classified as agricultural or non-agricultural. An agricultural tract is classified as a farm if its entire operation qualifies with at least $1,000 in sales or potential sales. All non-agricultural tracts and agricultural tracts with less than $1,000 in sales are classified as non-farms (Abreu et al., 2010). The JAS is used to quantify the number and type of farms that are not on the CML when producing Census estimates. During Census years, NASS increases the JAS sample size by adding segments from the Agricultural Coverage Evaluation Survey (ACES).

2.2 The Census and the Census Mail List

The Census utilizes a list-based frame, where the list contains both agricultural operations that are in the target population (i.e., farms) and agricultural operations that are not in the target population (i.e., non-farms) (Abreu et al., 2018). List building activities for the CML include updating list information from respondents to the previous Census of Agriculture and utilizing information from the National Agricultural Classification Surveys (NACS) that identify potential agricultural operations. Extensive efforts are directed towards developing a Census Mail List that includes all farms in the U.S (USDA National Agricultural Statistics Service, 2014, 2019).

To identify JAS records that are not on the CML, the names and addresses collected in the JAS are matched to the CML. Those names from the JAS that do not match a CML record are determined to be in the NML domain and a Census report form is sent to these records (USDA National Agricultural Statistics Service, 2019). Instructions on the Census report form direct any respondent who received duplicate forms to complete only one form and to mail all duplicate forms back together. Those who returned a CML and NML form are misclassified as NML and are removed from the NML domain (USDA National Agricultural Statistics Service, 2014, 2019). In the 2012 and 2017 Censuses, NASS applied a capture-recapture approach (Young et al., 2013, 2017) for producing estimates that are adjusted for undercoverage, non-response, and misclassification. The different adjustment weights are estimated for all responding Census records from logistic regression models fitted to JAS data and Census records linked to the JAS. The final published Census estimates are obtained by calibrating the model-based estimates to known commodity targets (Sartore et al., 2019).

3. Methodology

The existing Census estimation approach (Young et al., 2013, 2017) develops the estimation models based on the JAS/NML & CML matched dataset. Thus, CML records that are not captured by the JAS are not used for estimating model parameters. The proposed method utilizes all available JAS and Census records for estimating adjustment weights for Census farms. Estimation is conducted based on a capture-recapture approach similar to Alho (1990), but the likelihood from the new method includes parameters for the estimation of the probability of response. Instead of modeling a conditional capture, we are modeling the probability of capture in each survey and assume conditional independence to describe the probabilities of joint capture. The proposed method adds a penalty function in the log-likelihood for Census weights that are greater than 6 to avoid unrealistically high estimates at the county level. It is assumed that the Census/CML and the JAS records make up the first and the second draws from the population, respectively. To make the estimation feasible, each responding Census record must have JAS information. Similarly, Census information must be available for each farm record captured by the JAS. For this reason, data are shared between the Census and the JAS for some of the records. That is, the Census information is treated as JAS data for all Census records not included in the JAS sample, and the JAS information is used for non-responding and non-farm Census/NML records that are JAS farms. However, only responding Census records are used for producing estimates.

Unlike the existing Census model, the proposed approach assumes that misclassification can be ignored. However, the validity of our assumption needs further investigation. While variance estimation is not discussed in this paper, research has shown that parametric bootstrap or delete a group jackknife methods can be applied for estimating variances from the proposed method. In the existing approach, variance is estimated by using a combination of jackknife and bootstrap methodologies.

3.1 Individual Contribution of a Sample unit to the Likelihood

Let N be the number of farms in the population. Ignoring the fact that the JAS is based on a sample from the area frame, assume that each farm i is captured by the CML and the JAS with probabilities $\pi_{C,i}$ and $\pi_{J,i}$ respectively. That is, the capture events by the CML ($y_{C,i}$) and the area frame ($y_{J,i}$) have the following distributions.

$$y_{C,i} \sim \mathsf{Bernoulli}(\pi_{C,i})$$

$$y_{J,i} \sim \mathsf{Bernoulli}(\pi_{J,i})$$

Let r_i be an indicator variable for responding to the Census questionnaire. Response to the Census is assumed to have the following distribution:

$$r_i | y_{C,i} \sim \mathsf{Bernoulli}(\rho_i)$$

The likelihood contribution of the i-th sample unit is

$$L_i = \frac{\left\{\pi_{C,i}\rho_i^{r_i}(1-\rho_i)^{1-r_i}\right\}^{y_{C,i}} (\pi_{J,i})^{y_{J,i}}(1-\pi_{J,i})^{y_{C,i}(1-y_{J,i})}(1-\pi_{C,i})^{y_{J,i}(1-y_{C,i})}}{K_i^{y_{C,i}+y_{J,i}-y_{C,i}y_{J,i}}} \tag{1}$$

where K_i is the normalization constant:

$$K_i = \pi_{C,i}(1-\pi_{J,i})\rho_i + \pi_{J,i}(1-\pi_{C,i}) + \pi_{C,i}\pi_{J,i}\rho_i + \pi_{C,i}\pi_{J,i}(1-\rho_i).$$

Covariates are used to estimate $\pi_{C,i}$, $\pi_{J,i}$, and ρ_i as follows:

$$\pi_{C,i} = \frac{\exp(\mathbf{X_{c,i}}\beta^c)}{1+\exp(\mathbf{X_{c,i}}\beta^c)}$$

$$\pi_{J,i} = \frac{\exp(\mathbf{X_{j,i}}\beta^j)}{1+\exp(\mathbf{X_{j,i}}\beta^j)}$$

$$\rho_i = \frac{\exp(\mathbf{X_{r,i}}\beta^r)}{1+\exp(\mathbf{X_{r,i}}\beta^r)},$$

where, $\mathbf{X_{c,i}}$, $\mathbf{X_{j,i}}$, and $\mathbf{X_{r,i}}$ are respectively matrices of covariates for the Census, the JAS and the response probabilities, and $\boldsymbol{\beta}^c$, $\boldsymbol{\beta}^j$, and $\boldsymbol{\beta}^r$ are the corresponding vectors of coefficients.

3.2 Adjustment for JAS sampling probabilities

Because the JAS sample is drawn from the area frame, farm i in the population can be captured by the survey only if it is on the area frame, and then included in the JAS sample. So, the probability of capture by the JAS ($\pi_{J^a,i}$) is the product of the probability of capture by the area frame, and the probability of capture by the JAS given the record is on the area frame. Thus,

$$\pi_{J^a,i} = \pi_{A,i}\pi_{JA,i},$$

where, $\pi_{JA,i}$ is the probability of inclusion into the JAS sample, and $\pi_{A,i}$ is the probability of capture by the area frame. For each record, $\pi_{JA,i}$ is the reciprocal of sampling weights from the combined JAS and ACES sample, and $\pi_{A,i}$ is specified as a function of covariates as:

$$\pi_{A,i} = \frac{\exp(\boldsymbol{X_{a,i}}\boldsymbol{\beta}^a)}{1+\exp(\boldsymbol{X_{a,i}}\boldsymbol{\beta}^a)},$$

where, $\boldsymbol{X_{a,i}}$ is the covariate matrix and $\boldsymbol{\beta^a}$ is the corresponding vector of coefficients.

The likelihood is obtained by replacing $\pi_{J,i}$ by $\pi_{J^a,i}$ in Equation (1).

3.3 Penalized likelihood

High model-based weights can result in unrealistically high Census estimates at the county level. For this reason, weights are adjusted not to exceed 6 during the calibration process after model fitting is completed. To avoid extreme model-based weights, the proposed method penalizes weights that exceed 6 during the model fitting process. This is accomplished by adding a penalty function to the log-likelihood obtained from Equation (1). Let P_i be the reciprocal of the capture-recapture weight $\mathrm{DSE}_i = (\hat{\pi}_{C,i}\hat{\rho}_i)^{-1}$, which is used for producing Census estimates. The value of the penalty function, Pen_i, for record i is given by Equation 2.

$$\mathrm{Pen}_i = \frac{\log(\mathrm{P}_i)}{1+\exp(\lambda(\mathrm{P}_i - 1/6))}, \tag{2}$$

where, λ is a large positive number (e.g., $\lambda = 300$). The penalty function is close in value to the logarithm of P_i assuming that λ is big and $\mathrm{P}_i < \frac{1}{6}$. The individual contribution from record i to the log-likelihood is equal to $\ell_i = \log(L_i) + \mathrm{Pen}_i$.

3.4 The Estimation Process

Parameter estimates are obtained by maximizing the penalized log-likelihood

$$\ell = \sum_{i \in (\mathcal{C}\cap\mathcal{R})} \ell_i, \tag{3}$$

where, $\mathcal{C}\cap\mathcal{R}$ denotes the set of responding Census records.

The population size is estimated by

$$\hat{N} = \sum_{i \in \mathcal{C}\cap\mathcal{R}} (\hat{\pi}_{C,i}\hat{\rho}_i)^{-1}, \tag{4}$$

where $\hat{\pi}_{C,i}$ and $\hat{\rho}_i$ are respectively the estimated probabilities of coverage and response. Since the likelihood uses all Census and JAS data, the capture and response probabilities are obtained directly as fitted values for all $i \in \mathcal{C}\cap\mathcal{R}$ without requiring the computation of model-based predictions.

Similarly, land in farms and other quantities can be calculated as

$$\hat{Q} = \sum_{i \in \mathcal{C} \cap \mathcal{R}} Q_i (\hat{\pi}_{C,i} \hat{\rho}_i)^{-1} \tag{5}$$

where $\hat{Q}$ is the estimated value of the quantity of interest, and Q_i is the recorded Census value of the quantity from the i^{th} record.

4. Simulation study

Simulation studies were conducted to evaluate the performance of the proposed capture-recapture method before the penalty function is added. The simulations were conducted with and without adjusting for JAS sampling probabilities. The capture and response events and covariates are generated as follows for the case where JAS sampling probabilities are ignored. The results are similar to when the sampling probabilities are included.

For record i ($i = 1, 2, 3, ..., N$), let $Y_{C,i}$ & $Y_{J,i}$ be indicators for capture by the CML and the JAS respectively. Then,

$$Y_{C,i} \sim \mathsf{Bernoulli}(\pi_{C,i})$$
$$Y_{J,i} \sim \mathsf{Bernoulli}(\pi_{J,i}).$$

Where,

$$\pi_{C,i} = \frac{\exp(1.5 + X_{11i} + X_{12i})}{1 + \exp(1.5 + X_{11i} + X_{12i})},$$

$$\pi_{J,i} = \frac{\exp(-2.5 + 0.5X_{21i} + 0.5X_{22i})}{1 + \exp(-2.5 + 0.5X_{21i} + 0.5X_{22i})},$$

and, $X_{11i} = X_{21i} \sim \mathsf{Bernoulli}(0.3)$, $X_{12i} = X_{22i} \sim \mathsf{Bernoulli}(0.4)$.

Each farm on the CML is assumed to have a response probability that depends on covariates. The response status $R_i|Y_{C,i}$ is distributed as:

$$R_i|Y_{C,i} \sim \mathsf{Bernoulli}(\rho_i)$$

where,

$$\rho_i = \frac{\exp(0.5 + 2X_{ri})}{1 + \exp(0.5 + 2X_{ri})}, \quad X_{ri} \sim \mathsf{Bernoulli}(0.9).$$

Under this setting, about 88.5% of the farms in the population are captured by the CML, of which 89% are responding to the Census (i.e., about 79% of records in the population responded to the Census). About 11% of the farms in the population are captured by the JAS, and about 1% are captured by the JAS only (i.e., NML records). About 10% of the farms in the population are captured by both the CML and the JAS, and 9% of the farms are on the CML, responded to the Census, and are captured by the JAS. Note that all operations in the population are assumed to be farms.

The simulations were repeated 500 times, and starting values of 0 were used for all parameters. The number of farms is estimated by summing the capture-recapture weight

$DSE_i = (\pi_{C,i}\rho_i)^{-1}$ over responding CML records. In all of the population sizes considered, model based estimates of farm numbers are close to the true values. Figure 1 shows the frequency distribution of the estimated farm numbers for the case where the true population size is 1,000.

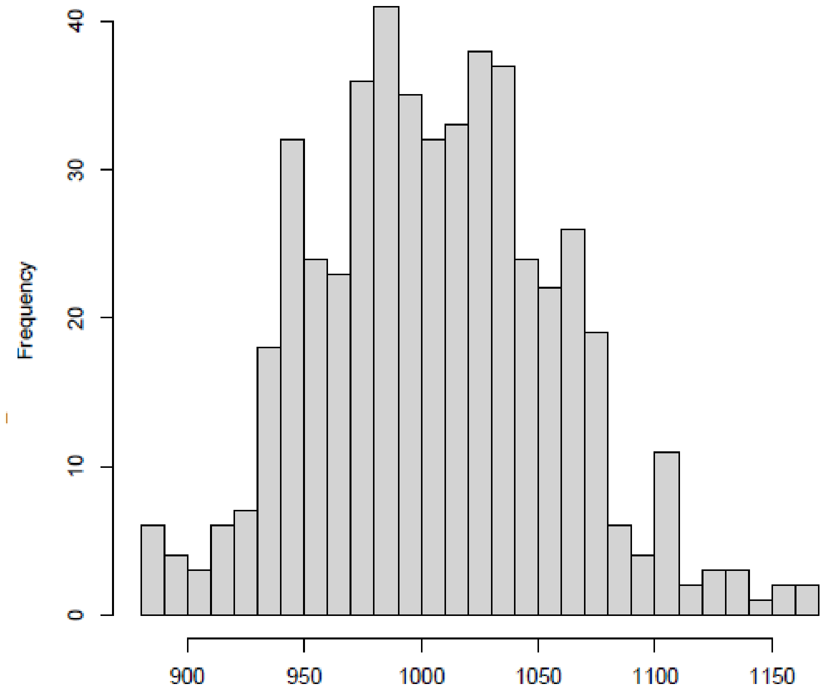

Figure 1: Frequency distribution of the number of farms estimates with true population size at 1,000 units.

5. Application to the 2017 Census Data

The proposed method is applied to the 2017 Census of Agriculture data for producing the number of farms and land in farms. As discussed in the previous sections, the final Census estimates are produced by summing the reciprocal of the product of the estimated CML capture probability ($\hat{\pi}_{C,i}$) and the probability of response ($\hat{\rho}_i$) over all Census farms. The capture probability, $\hat{\pi}_{C,i}$, accounts for undercoverage in the CML. The likelihood function includes the probability of capture by the JAS, but these probabilities are not used in the Census estimation. To produce the capture and the response probabilities (i.e., the final capture-recapture weight), all the Census farms, and records in the matched JAS-CML-NML data that have information from at least one of the three sources are used. Information is shared between the Census/NML and the JAS records when data are not available from one of the sources. In addition, it is assumed that the probability of a farm being counted as a non-farm is equal to the probability that a non-farm is counted as a farm in the Census.

Two approaches are used for fitting the models. In the first approach, a model is fitted for the entire U.S. and state-level estimates were produced from the model. The second approach involves producing state estimates from separate models fitted to a state or a combination of states. Model covariates are selected based on subject matter expert suggestions, and include variables representing demographic characteristics of operators as well as characteristics of farms.

The number of farms and land in farms estimates from the proposed model and the model-based estimates obtained in 2017 are compared to the published numbers by using percent relative differences (from the 2017 published number of farms and land in farms). Figure 2 shows percent relative differences of the number of farms, and Figure 3 shows percent relative differences of land in farms for a set of states and the U.S.

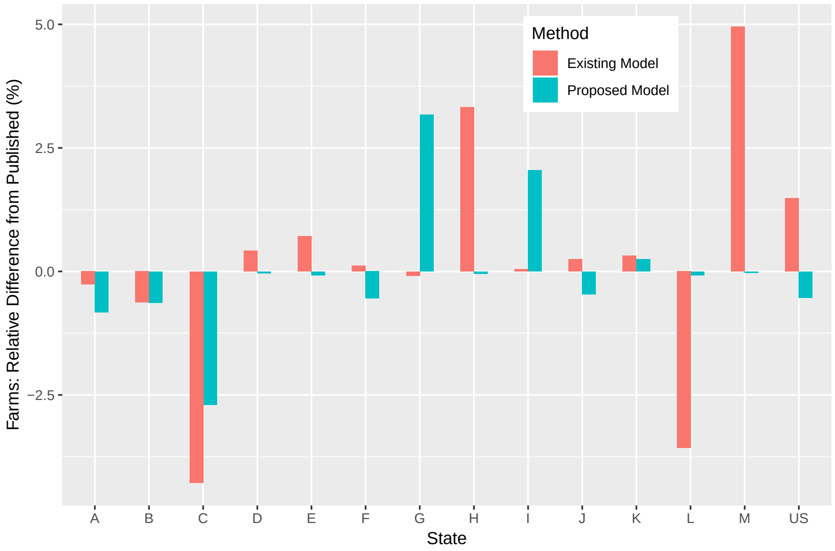

Figure 2: Percent relative differences of the number of farms estimates from the proposed and the existing models against the published estimates.

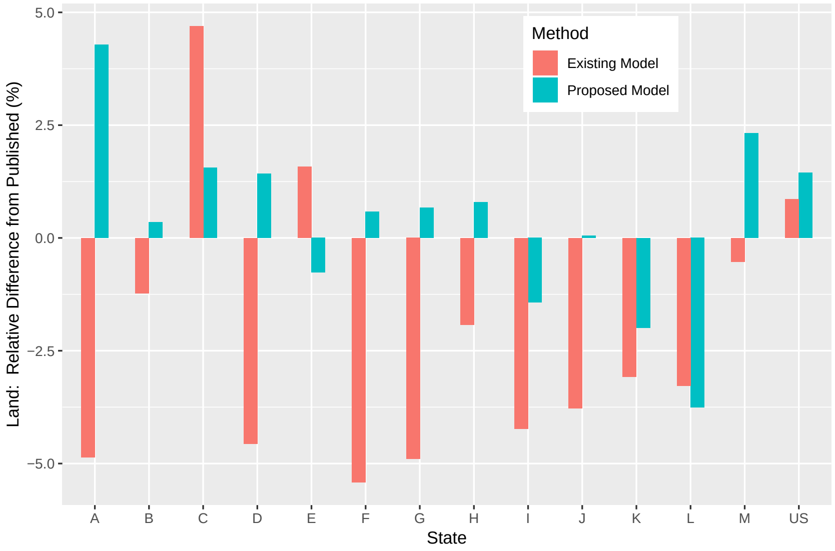

Figure 3: Percent relative differences of land in farms estimates from the proposed and the existing models against the published estimates.

For the majority of the states, estimates from the proposed model are closer to the published numbers of farms and land in farms compared to those from the existing model. The existing model performed better than the proposed model for the remaining states. We expect further improvement with the proposed model once we incorporate an adjustment for misclassification.

6. Conclusion

The proposed model can be considered as an extension of the methods discussed in Alho (1990). In fact, every Census and JAS farm is used for estimation of the model parameters governing the probabilities used for the computation of record-level Census weights, which will be finally calibrated and used to produce official statistics for the U.S. Census of Agriculture.

The application of the penalized log-likelihood proposed in Section 3.3 provides a solid foundation for stabilizing the computation of the optimal set of parameters. This also allows for a straightforward model fitting procedure that has been implemented and tested using SAS, R, and python. In comparison with the existing approach, the new method enables the estimation of Census weights without the need to perform predictions for records outside the JAS & CML matched dataset. In the proposed method, the fitted values of the probabilities for coverage and response to the Census questionnaire are automatically computed for every record in the CML (and also JAS) while parameters are being optimized.

The estimates computed at the U.S. and state levels are promising, although more work needs to be done to address challenges such as adjustment for misclassification. While the proposed method assumes that misclassification can be ignored, data from the 2012 & 2017 Census show substantial misclassification of farms in the Census. Thus, further research is needed to investigate suitable methods of accounting for misclassification. In addition, challenges in the estimation of variances and covariate selection need to be addressed before the new method can be implemented.

Acknowledgments

The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy. This research was supported in part by the intramural research program of the US Department of Agriculture, National Agricultural Statistics Service (NASS).

References

Abreu, D., Lawson, L., and Hickman, S. (2018). Assessment of a Review Process for the 2017 Census of Agriculture . In *In JSM Proceedings, Survey Research Methods Section. American Statistical Association.*

Abreu, D., McCarthy, J., and Colburn, L. (2010). Impact of the Screening Procedures of the June Area Survey on the Number of Farms Estimates. *Research and Development Division. RDD Research Report Number RDD-10-03. Washington, DC: USDA, National Agricultural Statistics Service*.

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Alho, J. (1994). Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics*, 10:245–256.

Alho, J., Mulry, M., Wurdeman, K., and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-systems estimation. *Journal of the American Statistical Association*, 88:1130–1136.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

Lamas, A., Abreu, D., Arroway, P., Lopiano, K., and Young, L. (2010). Modeling misclassification in the june area survey. *In JSM Proceedings, Survey Research Methods Section. American Statistical Association.*

Sartore, L., Toppin, K., Young, L., and Spiegelman, C. (2019). Developing integer calibration weights for census of agriculture. *Agricultural, Biological and Environmental Statistics*, 24(1):26–48.

USDA National Agricultural Statistics Service (2014). U.S. Census of Agriculture: United States Summary and State Data, Geographic Area Series.

USDA National Agricultural Statistics Service (2019). U.S. Census of Agriculture: United States Summary and State Data, Geographic Area Series.

Young, L., Lamas, A., and Abreu, D. (2017). The 2012 Census of Agriculture: a capture–recapture analysis. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):523–539.

Young, L., Lamas, A., Abreu, D., Wang, S., and Adrian, D. (2013). Statistical methodology for the 2012 U.S. Census of Agriculture. In *In the Proceeding 59th ISI World Statistics Congress*, pages 1063–1068.